48

SOME STUDIES ON AUTOMATIC SPEECH CODING AND RECOGNITION PROCEDURE

DWIJESH DUTTA MAJUMDAR AND ASOKE KUMAR DUTTA

ELECTRONICS DIVISION
INDIAN STATISTICAL INSTITUTE
CALCUTTA-35, INDIA.

(*Received June* 29, 1967; *Resubmitted April* 20, 1968)

(Plate 13)

ABSTRACT. The paper deals with the construction of a digital code pattern from a human speech input suitable for the use of computers from which the intelligence content can be extracted without any human intervention.

In the first part of this paper the major functions necessary in any pattern recognition system are noted, aspects of decision theory as pertinent to the present work are summarised, and a theoretical model for speech analysis by synthesis is described.

The second part of the paper describes the experimental part of the system, human speech analyser and matrix coder. Most of the circuitry are conventional, although some special circuits designed are for critical applications.

The underlying theme of the work is to design a frequency-quantized, continuous, short-term spectrum analysis technique capable of extracting statisically invariant properties of the human speech pattern to a binary quantized representation produced by the measurement apparatus and used as the input data to the categorizer portion of the recognition system.

INTRODUCTION

In recent years much attention is being given to the problem of recognition of speech by machine primarily for its potentialities towards a solution of the complicated problems of intelligence communication from man to machine. Though there has been considerable progress in the research and development side of the machine translation of languages, the subject of spoken word recognition and coding seem to be in its infancy. A system for speech recognition can be considered as an automaton which can extract the full inteligence content of speech and interpret the messages contained in it for decision-making and information-transferring purposes. The basic problem therefore reduces, following the development of the process of writing, into first forming a relatively compact set of codes distinguishing each'phoneme', i.e. the distinct elementary sounds and then analysing the ensemble of phonemes on the basis of linguistic knowledge and extracting the intelligence with the help of a suitable grammar. Such a code-system would be analogous to the short hand transcription in so far as the full intelligence content, the repre-

sentation of a particular speech sound (phoneme) by a particular code and the feasibility of handling it later (in a digital machine) are concerned.

The keen attention in the field of machine translation of languages has led to the considerable advancement in the analysis of words (assemblages of phonemes) in extracting intelligence. Compared to this the problem of forming usable, compact coding system directly from the analysis of speech acoustics is still in its infant stage, as a result, there exists a wide gap to be bridged for the achievement of recognition. The objective of this paper is to describe a "Spoken-word Recognition System" being designed in this Laboratory, along with the theoretical background, the format of the generated coding system, and statistical classificatory method for the recognition of the codes. A very brief review of relevant works in this field is given at the outset.

A BRIEF HISTORICAL PERSPECTIVE

The first automatic speech recogniser was designed by Jean Drefus Graf (1950), and after that has been dealt with by several research workers such as Davis *et al* (1952), Dudley and Balashek (1958), Wiren and Stubes (1956), Roman Jacobson (1952), Jacobson, *et al* (1956), Fry (1956), Denso and Fry (1959) etc., but the use of computers in speech recognition problem was first made by Forgie and Forgie (1959), and could recognise the ten vowels with 93% accuracy. The study of Peter Denes and Matheww (1960) showed that the computers provide considerable advantage for solving many of the problems encountered in speech research. Forgies (1962) made a significant computer study to recognise the fricatives /*f*/ and /*θ*/, as in fie and thigh, frill and thrill, ruth and myth etc. For the final fricatives the computers was as good as a human listener but for initial fricatives human listener was much better.

The early efforts of designing speech recognition machines depended almost exclusively on accoustical information of the input signal with little or no emphasis on the linguistic constraints. Extensive researches on accoustic cues began revealing alarming spread of the acoustic parameters for different speakers, different moods and for differing stress and intonation. The overlapping of the features began suggesting that some other criteria must be found out to resolve confusion and ambiguity. Quite naturally, attention was fixed on grammatical and syntactic contents of speech. The domain of research has been broadened even to the territories of psycho-linguists, speech and learning pathologists, and pure linguists (Lindgren 1965). Meanwhile work on actually building an automata which will recognise natural speech seems to be somewhat in abeyance. The present work presents some theoretical studies (Dutta Majumdar and Dutta, 1966) and an experimental model, the detailed results could not be given here to keep the size of the paper within a reasonable limit.

GENERAL PROBLEM OF AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is essentially a problem of mapping a continuous function into a set of discrete symbols. The input speech waves provide the acoustical data which is analysed and then transformed into a discrete set (figure 1). In generation and perception of speech a human being utilises prior linguistic know-

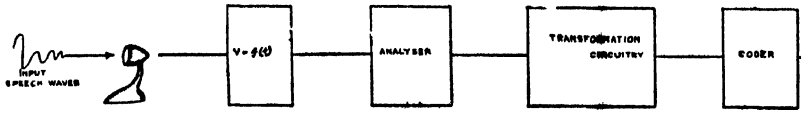


Figure 1: Generalised block diagram of automatic speech coding problem.

ledge and it is argued, therefore, that the acoustical data does not contain the complete information for the interpretation of speech. The total information may be dealt with in two different ways. The first is to generate the various details of speech on the basis of stored information for comparison, with the input speech wave and obtain a best match; this is known as the process of *analysis by speech simulation*. In the second, various aspect of the input waves are compared with stored elements and distributions; this process is known as *dynamic speech analysis*. If we symbolise the transformation from speech production (P) to discrete symbols (S) by $P \rightarrow S$ and the transformation from speech production (P) to speech acoustics (A) by $P \rightarrow A$ then the first process may be symbolised by $A \rightarrow P \rightarrow S$ and the second process by $A \rightarrow S$ (figure 2). Paterson (1961) found it more reasonable to follow the direct route $A \rightarrow S$ for the purpose of automatic speech recognition.

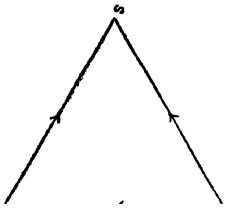


Figure 2: The transformation scheme.

Whereas the "motor theory of perception" forwarded by A.M. Libberman and his colleagues (1962) (1964) at Haskins Laboratory suggests the indirect process $A \rightarrow P \rightarrow S$. It may be observed, however, that there has been considerable controversy about the motor theory. According to Gunnur Fant (1964) speaking ability is not a necessary requirement but it enters as a conditioning factor. The model of speech production and perception by a human being shown in Fig. 3 was proposed by Fant. The motor and sensory centres become more and more

involved as one moves to the right. In the Haskins model the criteria of recognition is established by an association of the motor commands with the percieved auditory patterns. After a correct hit the decoding can be achieved by either of the branches KFE or CDE.

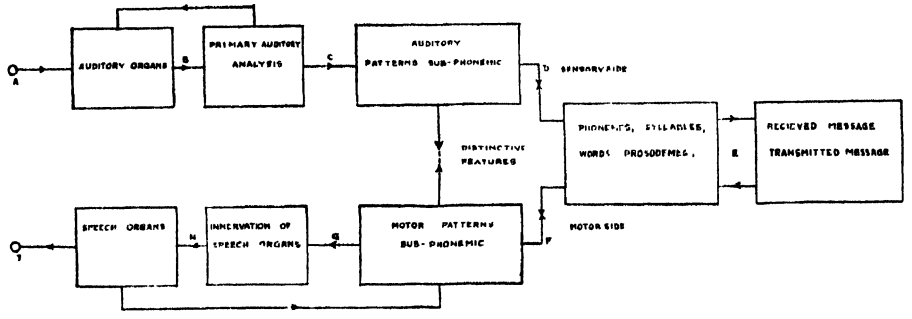


Figure 3: A hypothetical model of brain functions in speech production and perception

Stevenson and Halle (1958) postulated that in perceving speech, listener's brain first generates patterns with the help of certain previously learned rules and compares these patterns with the patterns obtained from the preliminary analysis of the incoming acoustical signals. He suggested that these "generative rules" are largely identical to rules used in speech production. The model based on their postulates shows that the course followed is an inter-connected complex of the two processes $A \rightarrow S$ and $A \rightarrow P \rightarrow S$ described by Peterson and Gordon (1961).

THEORETICAL BASIS FOR THE RECOGNITION PROCEDURE

The acoustical analysis of the signal space results in a measure $F(t)$, where $F(t), f_1(t), f_2(t) \ldots f_n(t)$ each being a binary valued time-function of the acoustical parameters. Each of $f(t)$ is represeted by a row of the code matrix. Owing to the varibility of the speaker and the variations associated with physiological processes in different utterances even by the same individual of the same word, there will be some variation in the measurement F. Each component may therefore, be visualized as a stochastic process with distribution being taken as normal. Thus the measurement space M constitutes a set of F. The whole recognition problem is based on the pre-supposition that it is possible to have a suitable classificatory method for grouping the elements of the measurement space corresponding to different syllables.

The measurement space M may be considered as an n-dimensional space of functions $f_1, f_2, \ldots, f_n$ as axes of reference. A measurement F represents a point in this space M. A syllable represents a cluster of points. The vocabulary of the machine is taken to be finite with N number of different syllables i.e. N different cluster of points in the measurement space. The problem, therefore, *is*

first to divide the space M into N regions $R_1, R_2, \ldots R_n$ such that there is minimum possible overlapping and then to assign an unknown individual to its proper group.

Let $P_1, P_2, \ldots, P_n$ be the proportions of the individuals of the N groups of mixed population where $\sum_{r=1}^{N} P_r = 1$. Also let $\phi_r(f_1, f_2, \ldots f_n/\theta)$; ϕ_r/f_θ in abbreviated form represent the probability density of the rth group, θ_r being the parameter entering in the probability desnity. The probability of a wrong classification is given by

$$\alpha = 1 - \sum_{r=1}^{N} P_r \int_{R_r} \phi_r dv \qquad \ldots (1)$$

The error will be minimum when $\sum_{r=1}^{N} P_r \phi_r dv$ is minimum. By generalised Neyman Pearson's Lemma such best possible regions in space are defined by (Rao, 1952)

$$R_r : p_r \phi_r \geqslant p_s \phi_s \text{ where } S \neq r$$

and

$$r, s = [1, 2, \ldots, N]$$

Neyman and Pearson has shown that the boundary separating those regions in space are defined by surfaces of constant likelihood ratios (Rao 1952). If the probability densities are multivariate normals with dispersion matrix λ_{ij} and the mean value μ_{ir} where r, j are variate indices and r is the group index, then the surfaces of constant likelihood ratios are given in terms of the discriminant function introduced by Fisher (Rao 1952)by

$$\sum_j \{ \sum_i \lambda_{ij} d_i \} f_j = \text{constant.} \qquad (2)$$

where

$$d_i = \mu_{ir} - \mu_{is} \quad r \neq s$$

$$\lambda^{ij} = \text{the reciprocal of } \lambda_{ij}$$

$$i, j = [1, 2, \ldots, N-1]$$

$$r, s = [1, 2, \ldots, N]$$

The probability density of the multivariate normal population is given by

$$f_r = \text{Const.}\ \lambda e^{-\frac{1}{2}\{\sum \sum \lambda^{ij}(f_i - \mu_{ir})(f_j - \mu_{jr})\}} \qquad \ldots (3)$$

The n-dimensional space M is thus divided into the regions R by the hyperplanes whose equ. is (2) for a suitably determined constant.

These surfaces can also be defined in terms of linear discriminant scores determined by the constants for the group only :

$$L_r = \sum_j \{(\Sigma\lambda^{ij}\mu_{ir})f_j\} - \tfrac{1}{2}\Sigma\ \Sigma\lambda^{ij}\mu_{ir}\mu_{jr} \tag{4}$$

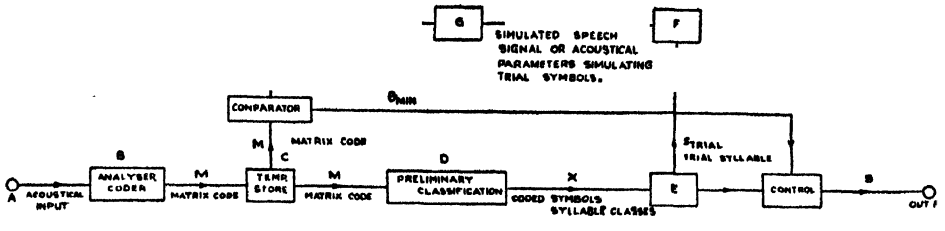
$$r = [1, 2, .., N] \tag{5}$$

A constant likelihood ratio corresponds to a constant difference in the discriminant score L. If the a priori probabilities are $\pi_1, \pi_2, .., \pi_n$ for the N groups then an individual is assigned to the group for which $L_r + \log \pi_r$ is a maximum.

A MODEL FOR SPOKEN-WORD RECOGNITION

The present authors suggest the model schematically shown in figure 4, which is the logical development of figures 1, 2, and 3, based on the model of Halle and Stevenson but with the basic modification in approach that the present analysis is based on syllables as the elementary or fundamental linguistic "atoms" constituting the phonemes whereas the other one is based on phoneme. In appendix I, table 1 is given the list of english phonemic elements, and in table 2 is given their distinctive features (choice between two opposites) pattern. Through sets of systematic comparisons and a series of two-choice selections binary minimal contrasts between word elements are isolated. As for example vowel /0/ in table 2, is vocalic, nonconsonant, compact, grave and flat, whereas the vowel /a/ has all the above features and differs only in flat/plain feature.

The incoming speech signal undergoes acoustical analysis and sampling and is then transformed into a matrix code in the Analyser-Coder unit B. This unit has been designed and constructed by us in our laboratory and is described in detail with some experimental results in a following section. This matrix code is temporarily stored in block C, which may or may not be a part of the on-line computing system to which this matrix code is fed. The computing system with concurrent programming facility is shown in figure 4, as comprising of different blocks to facilitate explanation. The adaptive algorithm as explained in the previous section programmed into D classifies the incoming codes (M) into



igure 4 : The block diagram of the suggested model for spoken word recognition.

classes of same or phonetically very alike syllables. First a previously classified representative set of core syllables are fed into the machine. This constitutes the "previous knowledge" of the machine. The adaptive alogorithm makes it possible to modify this previous knowledge whenever a correct decision is made. This is done by constantly re-computing f_r, the probability density of the population of each class whenever a new symbol is correctly recognised. Thus the surfaces of constant likelihood ratios separating the regions corresponding to different symbols undergoes constant modification on the basis of each correct new score. The a- priori probabilities also undergoe such modification. This procedure may be called the "learning process" of the machine as the "basic knowledge" is constaltly improved by the "additional knowledge" gained through each correct classification.

At D also a tentative decision for the transformation $M \rightarrow X$ is made. Where X is the class of symbol of measurement space M, partitioned according to criterion described earlier. This decision is made by computing $L_r + \log \pi_r$ (equation 5) and finally the class X for which this expression is maximum. There is a possibility that, for sample lying close to the separating boundry this maximum may not always be welldefined resulting in ambiguity and/or incorrect classification. To resolve this the decision is further subjected to examination on the basis of extra-acoustical information. In the next step a trial transformation to symbols $X \rightarrow S$ is made at block E of figure 4. This transformation uses phonological and linguistic rules and also examines the contextual basis of the trial made. The output from this S trial is passed on to F where the transformation to speech production $S \rightarrow P$ is made using rules simulating speech synthesis from abstract phonetic symbols and its output P_{trial} simulates speech signal patterns or some patterns which simulates those derived from speech signals. This P_{trial} is analysed and coded at G whose principle of action is closely analogous to that of B, the basic difference being that whereas in B the actual acoustic (analog) signal is treated, in G synthetically produced digital speech signal is treated to produce M_{trial}. This M_{trial} is fed into a comparator unit. The comparator evaluates the discrepancy between M_{trial} and M from the temporary store C. The symbol S for which this discrepancy or error is minimum is taken to be the correct syllable symbol and a control is actuated so that the symbol S is passed to the output. If the minimum value of θ is above the previously set value for a correct recognition, a reference is made at a lower level of identification and the process is repeated. This system allows for nonsense syllables (nonsense in the context of the limited vocabulary of the machine to be branded as unknown and further reduces the risk of wrong classification.

DESIGN OF SPEECH ANALYSER AND MATRIX CODER

Classification of speech Sounds : Peterson and Gordon (1961) suggested the following four classification of acoustical speech sounds :

space alone. But for the identification of the initial or final consonants the formant movement in the real time should be noted as a marked change in the on-glide or off-glide due to vowel-consonant interaction. The band spread of the formants and the variation of the spread with time are the acoustical characteristics necessary for the identification by the inherent sonority features of vocalic/non-vocalic, compact/ difuse, tense/lax and nasal/oral phonemes (Belar and Olson 1962). Another parameter of particular importance in distinguishing the sibilant consonants specially $|S|$, $|\int|$ and $|Z|$ is the noise component of the spectrum. Barozinskl's analysis (1934) shows a main noise range of 6 to 9.5kcs for $|S|$ and 5-12kcs for $|j|$.

Though the average speech power is of very little or no importance for recognition purposes instantaneous speech power plays a decisive role in distinguishing some inherent distinctive features. The rapid fluctuation of the instantaneous speech power characterises certain difficult liquid phonemes such as $|l|$ and $|r|$. These trills are associated with rapid fluctuation of the noise spectrum.

Work on CNC (Consonant phoneme—Vowel Nucleus—Consonant phoneme) syllables by Lehiste and Peterson (1961) shows the importance of the duration of the initial and final transitions with respect to the target duration for distinguishing certain vowel sounds. The four short vowels $|I\varepsilon\text{ə}U|$ have a relatively long off-glide and a corresponding shorter target; the long vowels have a relatively shorter off-glide and longer target. The vowels $|e^I|$, $|0^0|$ and $|3|$ called glides, can not be classified as dipthongs because of the difficulty of resolving them into a sequence of two sounds. Their main characteristic is that the target duration is comparable to either the on-glide or the off-glide [fiigure 5]. The dipthongs aI, aU and $3I$ are characteriised by two distinct target positions. The first target is usually longer than the second target and the transition between the targets are longer than either target.

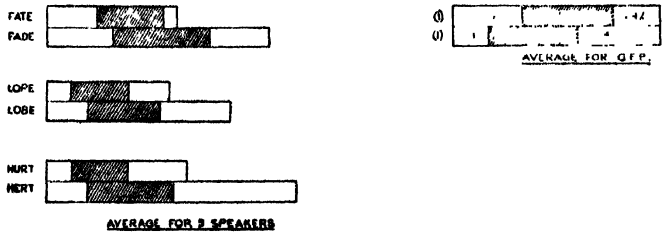


Figure 5 : Percentage time of target duration.

On the basis of the above discussion the following acoustical parameters need be recorded for the generation of a coding system for speech recognition :

A) Resonance bands for formant locations :

1) Band (i) —250 c/s—350 c/s

2) Band (ii) —350 c/s—500 c/s
3) Band (iii) —500 c/s—650 c/s
4) Band (iv) —650 c/s—900 c/s
5) Band I 800 c/s—1050 c/s
6) Band II 1050 c/s—1350 c/s
7) Band III 1600 c/s—3000 c/s
8) Band (S_1)4500 c/s—above
9) Band (S_2) 10,000 c/s—above (provisional)

The choice of the bands 1—7 have been based on the works of Peterson and Barney (1952) (vide table 1) and the sibilant bands have been based on Barozinski's analysis as discussed earlier

B) Inflection : Code 00—No voice
Code 01—Rising voice
Code 10—Falling voice
Code 11—Steady voice

C) Intensity : Code 00—0db—+6db
Code 01—+db—+12db
Code 10—+12 db—+18db
Code 11—+18 db—above

D) Time duration : Code 00— 56ms
Code 01— 56ms—122ms
Code 10—122ms—250ms
Code 11— 250 ms—above.

E) Trill or Roughness Measure

Table 1 Taken from Peterson Barney analysis of ten vowels :Average for 33 men and 28 women are only given

Formant frequencies c/s										
F_1	Men	270	390	530	660	730	570	300	440	640	490
	Women	310	430	610	860	850	590	370	470	760	500
F_2	Men	2290	1990	1840	1720	1090	840	870	1020	1190	1350
	Women	2790	2480	2330	2050	1220	920	950	1160	1400	1040

SAMPLING SCHEME AND CODE FORMAT

The measurement space of the system consists of a set of acoustical parameters (listed in the precoding section) which are continuous functions of time. Except for the intensity and time duration it is sufficient to examine whether the measures of the functions cross certain thereshhold values. Such measurements make the reduction of the set of continuous functions to a corresponding set of binary valued time functions possible. Only a suitable sampling scheme is to be decided upon for the purpose of coding these measurements. Sampling at regular intervals has been suggested by many authors and is a sampling method more common than the method of sampling on detected change in the functions. King and Tunis (1966) have suggested in contradiction to the views of Clapper (1964), Olson and Belar (1962) that the sampling at regular intervals is a more efficient method. The efficiency of any sampling scheme lies in reduction of the volume of data without significant loss in the information. The more accurately the original function can be reconstructed without increasing the volume of data the better is the sampling scheme. Obviously if a sample of the binary valued time function be made whenever there is a change and the duration of the period of no change is noted, the whole function can be reconstructed with an accuracy which is wholly dependent on the accuracy of the time measurement (the sampling dead time being usually neglible). In fact it is somewhat like writing down the whole function in a different form and is not a sampling in the strict sense. Poorer score in sample-on change scheme to a regular sampling scheme at 30m sec. interval as observed by King and Tunis may be due to the fact that there had been no measure of time without which former scheme loses much of its significance. Moreover, how far the simulation of this scheme from the data of regular sampling is representative of the actual scheme seems doubtful. Since the actual problem envisages a dynamic analysis the changes are much more significant than is apparent, the accuracy of depicting the changes in the time domain is therefore very important. The sample on-change emphasises this point which is lost when a sampling is made at regular intervals. This loss is not retrievable in a simulation experiment as envisaged in the experiment of King and Tunis.

The sample-on-change scheme therefore has been accepted in the design of the Matrix coder An experiment to assess the efficiency of this scheme in comparison with sampling at different regular intervals can be made by simulation on a digital computer. Since the function can be reconstructed with a high degree of accuracy from the proposed scheme the simulation should not introduce any appreciable error.

The code format of the matrix coder is a matrix of fixed rows and variable columns as illustrated in figure 6. The output of the filterbands will encode the format position as well as the presence or absence of sibilant noise. The matrix

element is coded as one of the output of the corresponding band filter exceeds a present threshhold value, otherwise the element is coded as zero. The consecutive columns keep track of any movement of the formant positions. The inflection code consists of two rows RV and FV corresponding to rising voice or falling voice. If both RV and FV are unity the interpretation is a steady voice and if both are zero, the interpretation is no voice. The interpretation of intensity, time and roughness codes are given in the previous section. A provision of 16 column representation has been kept though the code length is not expected to exceed 10 columns for the selected syllable vocabulary.

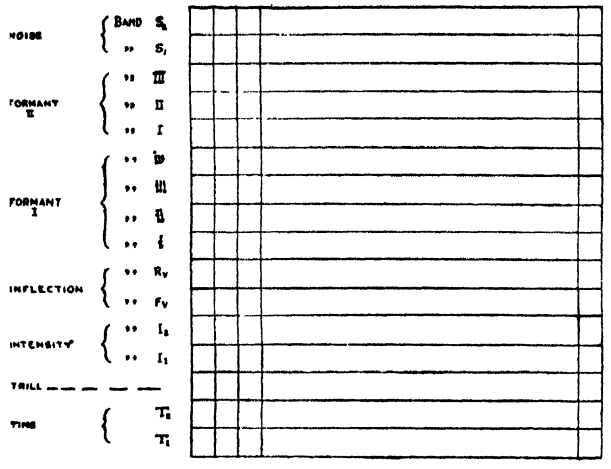


Figure 6 : Matrix coder with fixed rows and variable columns.

OPERATIONAL DESCRIPTION OF THE ANALYSER-CODER UNIT

A schematic diagram is shown in figure 7. The input microphone produces an electrical analogue of the acoustical waveform of speech which is amplified in the pre-amplifier with AGC control. The output of the pre-amplifier is thus amplitude normalized and is kept at an average of 900 mv peak to peak. This compressed waveform is fed into four band equalizer amplifier F_0, F_1, F_2 and S corresponding respectively to fundamental voice frequency band, first formant, second formant and noise band. The AGC voltage and the monitoring peaks of F_0 the speech waveform is obtained from the pre-amplifier as by-products. The AGC voltage is fed into roughness detection circuits and Instnsity Digitiser. The F_0 is fed into the Inflection Detection circuits. A timing diagram of the Analyser-Coder is shown in figure 8.

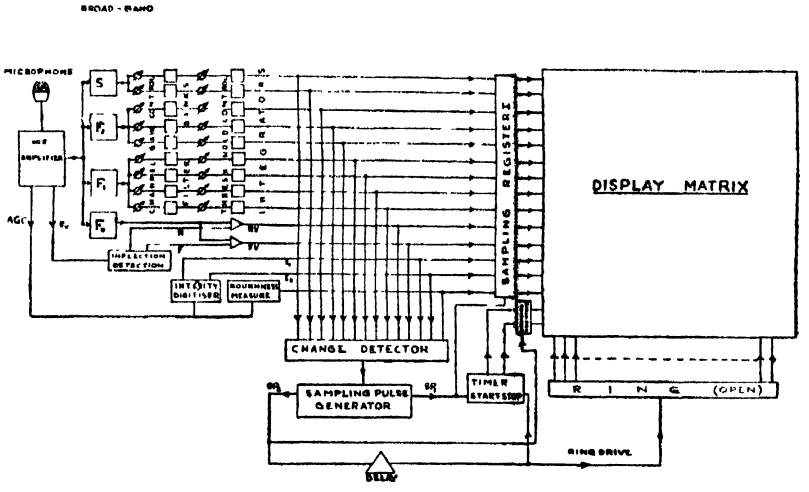


Figure 7 : Schematic diagram of the analyser-coder unit.

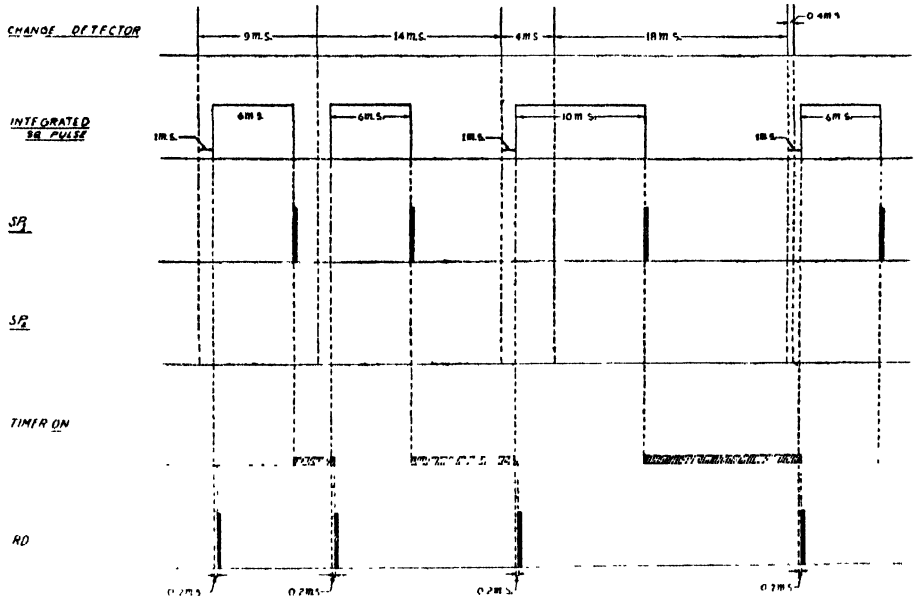


Figure 8 : Timing diagram in the analyser-coder unit.

The broad band equaliser amplifier normalises the general 6db/octave fall in higher voice frequencies. The output of each amplifier drives the corresponding filter band through individual Channel gain control circuits. The filter bands incorporate circuits with symmetrical response curves. The outputs of the individual Filter Bands are A.C. coupled to Integrator circuits provided with indivi-

dual threshhold control (figure 9). The output of the Integrator remains in the ground state unless the input reaches the threshold level and until the integrated value of the input reaches a minimum value. The output then switches to the 'ONE' state and ignore the absence of the input for very short intervals. The output switches back to ground state only after the input falls below the minimum level for a certain time.

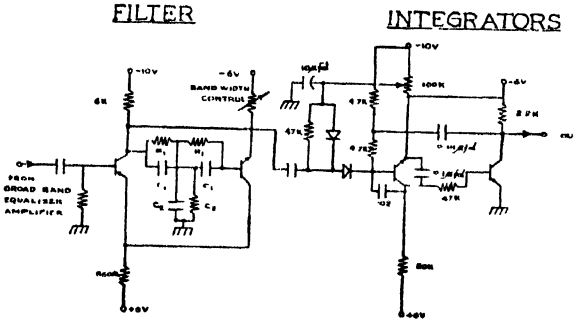


Figure 9 : Filter-Integrator circuit.

The spikes F_v corresponding to the peaks of the speech wave are obtained from the pre-amplifier for the detection of the inflection. F_v is integrated to give D.C. voltage which is proportional to the frequency of the spikes. Thus this D.C. voltage has $+ve$ or $-ve$ slope as the frequency increases or decreases. This slope is detected and is used to drive either of the two different transistors 'ON' depending on the sign of the slope. A final interlocking of the states of these transistors with the integrated output from the Broad-band Equaliser Amplifer F_0 provides the needed codes FV and RV (figure 10).

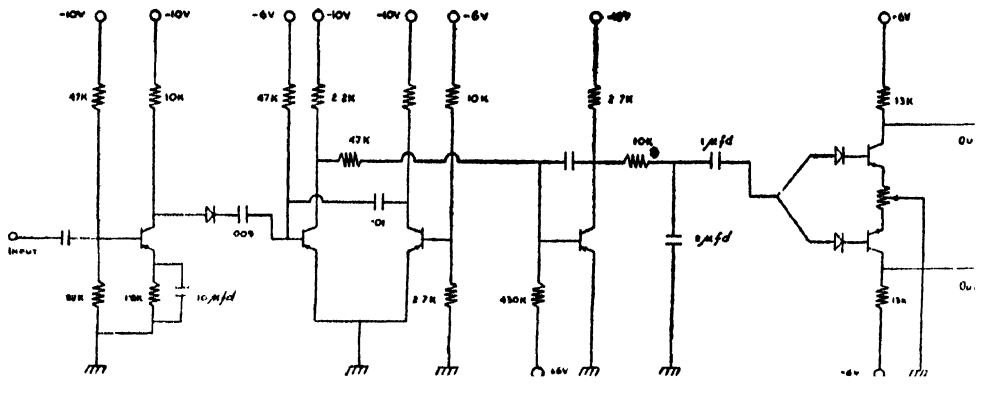


Figure 10 : Inflection coding circuit.

The AGC voltage from the Pre-amplifier is used for the detection of roughness and for measurement of intesity. The sharp changes in the AGC voltage correspond to the rapid fluctuations of the speech power. These short bursts are

filtered out and integrated in a biased integrator circuit very similar to that employed in the integrator circuits employed in the band integrators. For the purpose of measuring and coding intensity the AGC voltage is smoothed, amplified and compared with fixed reference voltage and on the basis of comparison two transistor switches are made to operate resulting in the required code. Speech envelopes (upper traces) and the corresponding detected and coded roughness (lower traces) for some typical words are shown in the photographic plate 13. Two sweeps are necessary to depict the traces of the word "Around", which is quite a long one.

Logarithmic scale for time measurement has been chosen for better resolution in short range and for covering the useful range without increasing the number of codes. The exponential decay of the voltage across a capacitor is fed into an analog-digital converter to form the time codes T_1 and T_2 at the output of two transistors (figure 11). The first transistor is turned on at about 56 ms after start signal, the second transistor remaining off, so that T_1 is high and T_2 is low resulting in the code 01. After about 122 ms the second transistor is turned on and the first one is turned off and the code 10 is generated. Finally at about 250 ms both the transistors are turned on and the code is 11.

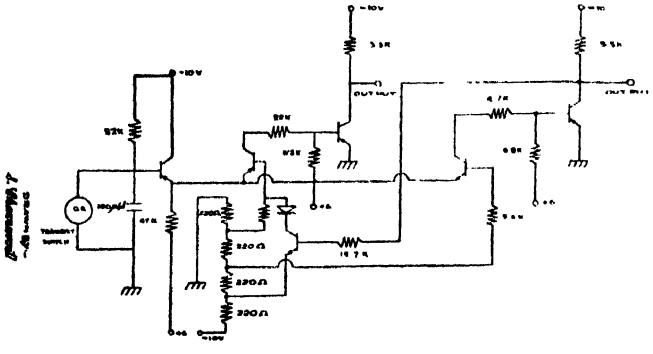


Figure 11 : Exponential time code generator circuit.

The Matrix display units employ incandescent lamps driven by silicon-controlled rectifiers. These SCR's are triggered by the ANDing of the output of sampling Registers with the outputs of ring counters.

Certain important and interesting circuits as have been used in the above units are given in this section. And at this stage it will be convenient to examine the sequence of operation of the whole instrument.

The output of all the information channels except the duration codes are fed into a change Detector which is notihng but a transient detector that produces about positive going spike for each change. These spikes are first quantised, integrated and then shaped into a binary value square pulses which remains "on"

for a specified period of time. The time lag between the occurrance of a spike and the generation of the square pulse is about 1ms. Any other spike coming within this time is not resolved and completely neglected.If any further change in any of the information channel occurs outside the resolution time but while the square pulse is 'ON' the pulse is just stretched. As sampling of the information codes are done by the trailing edge of this square pulse all the changes occurring during the on state of this pulse except the last change is not entered into the matrix. This is done, to reduce confusion by limiting data and to make the code compact. This resolution time can be varied to obtain the maximum recognition score. The differentiated output of this trailing edge generates the sampling pulse SP_1. The SP_1 samples the sampling register I, and also starts the timer. The positive edge of the square pulse is differentiated to get the sampling pulse SP2. This pulse samples the sampling register II. After delay of about 0.2ms a ring drive pulse advances the ring and stops the timer.

CONCLUSION

An attempt has been made to place the present state of the automatic speech recognition problem both from theoretical and experimental point of view. The general problem as it appears from the model of the recognition procedure presented in this paper may be regarded as being composed of two major parts : (1) primary recognition based solely on the sound shapes of the acoustic signal; and (2) a secondary recognition of the linguistic (grammatical and syntactic) content based on the phonemic output of the primary recognition level. These two major parts would undoubtedly be implemented in a machine in many complex hierarchies of procedures and decision strategies.

The present scheme of the machine is of course that of an experimental model, to make a deeper understanding of the inherent complicated problems possible. In the final design, it may be necessary to incorporate the faculties of the ordinary listener-knowledge of the meanings of utterances, rules of grammar, feelings of phonological probabilities, vast stores of general knowledge organized and codified in some form as associative system. The incorporation of such faculties in automata will depend on a deepgoing investigation and quantification of the dynamic functions of the central nervous system, and is an aspiration for the future researchers.

ACKNOWLEDGEMENT

The authors wish to acknowledge with thanks Professor S. K. Mitra, Head, Computer Development and Research Division, Indian Statistical Institute and Professor P. C. Mahalanobis, F.R.S., Prof. C. R. Rao, F.R.S., Director R.T.S. for their kind interest in the work, and Sarbasree S. Basu, R. Ganguli, J. Das and A. R. Das Gupta for their help in constructing the electronic circuits.

APPENDIX 1

Table 1

English phonetics

phonetic symbol	key word	phonetic symbol	key word
simple vowels		plosives	
I	f*i*t	b	*b*ad
i	f*ee*t	d	*d*ive
ɛ	l*e*t	g	*g*ive
æ	b*a*t	P	*p*ot
A	b*u*t	t	*t*oy
a	n*o*t	K	*c*at
∂	l*aw*		
		nasal consonants	
U	b*oo*k	m	ma*y*
u	b*oo*t	n	*n*ow
3	b*ir*d	η	si*ng*
∧	B*er*t		
complex vowels		fricatives	
ɔ	p*ai*n	z	*z*ero
O	g*o*	3	vi*s*ion
aU	h*ou*se	V	*v*ery
Al	i*c*e	δ	*th*at
∂I	b*oy*	h	*h*at
IU	f*ew*	f	*fl*at
		θ	*th*ing
		ʃ	*sh*ed
		s	*s*at
semivowels and liquids		affricatives	
j	*y*ou	tʃ	*ch*urch
W	*w*e	d3	*j*udge
l	*l*ate		
r	*r*ate		

Table 2
(Distinctive-features pattern of English phonemes)

	0	a	e	u	ə	i	l	ŋ	ʃ	ʧ	k	ʒ	ʤ	g	m	f	p	v	b	n	s	θ	t	z	ð	d	h	≠	≠
1. vocalic/ nonvocalic	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	9	−	−
2. consonantal non-cons.	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	
3. comp./diff	+	+	+	−	−	−	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	
4. grave/acute	+	+	−	+	+	−								+	+	+	+	+	−	−	−	−	−	−	−	−			
5. flat/plain	+	−		+	−																								
6. nasal/oral								+	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−			
7. tense/lax									+	+	+	−	−	−		+	+	−	−		±	+	+	−	−	−	+	−	
8. continuant/ interrupted									+	−	−	+	−	−		+	−	+	−		+	+	−	+	+	−			
9. strident/ mellow									+	+	−	+	+	−							+	−	−	+	−	−			

REFERENCES

Barezinski, L., 1934, *Winer Med. Wshr.*, 44.

Bolar, H., and Oloson, H, F. (1962), *IRE Trans. on Audio Au*—10, 11–17.

Clapper, G. L., 1964, *IBM Technical Report* (internal).

Davis, K. H. *et al*, 1960, *J.A.S.A.*, **32**.

Donse, P., and Fry, D. B., 1959, *J. Brit. IRE*, **19**

Denes, P. and Mathews, M. V., 1960, *J.A.S.A.*, **32**.

Dreyfus Graf. J., 1950, *J.A.S.A.*, **22**

Dudley, H. and Balshak, S., 1958, *J.A.S.A.*, **30**

Dutta Majumdor, D. and Dutta, A. K., 1966, *Proc. Symp. Control and Computation*, 9-10.

Fant, G., 1964, *AFGRL*, Boston

Forgie, S. W. and Forgie, C. D., 1959, *J.A.S.A.*, **31**

Forgie, S. W. and Forgie, C. D., 1962, *Fourth Int. Congress on Acoustics.*

Halle, M. and Stavens, K., 1962, *IRE Trans. on Information Theory*, Vol. II-8(2),

Jacobson, R., Fant, C. G. N. and Halle, M., 1952, *MIT Acountic Lab. Report.*

Jacobson, R. and Halle, M., 1956, *Fundamentals of Languages*, 8, Gravenhage, Netherlands Nouton & Co.

King, J. H., and Tunis, C. J., 1966, *IBM J. Res. & Dev.*, **10**(1)

Lehiste, I. and Peterson, E. G., 1961, *Communication Sciences Lab. Report, Univ. of Mich*, Ann Arbor.

Libberman, A. M., 1962, *Proc. Speech Comm. Seminer*, Stockholm.

Libberman, A. M., 1964, *AFCRL*, Boston.

Lindgren, Nilo, 1965, *IEEE Spectrum*,

Lindgren Nilo, 1965 *IEEE Spectrum*,

Peterson, G. E. and Brney, H. L., 1952, *J.A.S.A.*, **24**

Peterson, E., Gordon, 1961, *Language and Speech*, **4**(4),

Rao, C. R., 1952, *Advanced Statistical Methods in Research*, John Wiley and Sons, 351-355.

Wiren, J. and Stubbs, H. L., 1956, *J.A.S.A.*, **28**